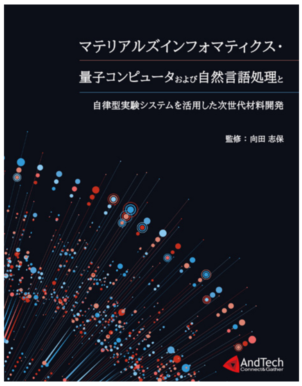
This booklet is a copy of the original chapter "Artificial intelligence for curation of information and knowledge acquisition" authored by Christopher L. Farrow, PhD and Alexandre Chabot-Leclerc, PhD of Enthought from the book "Next-generation Materials Development Using Materials Informatics, Quantum Computers, Natural Language Processing, and Autonomous Experimental Systems."

You are invited to go to <u>enthought.jp/mi-book</u> to download the chapter in both English and Japanese.



Enthought

Artificial intelligence for curation of information and knowledge acquisition

ABSTRACT

As competition in new material development intensifies, the importance of knowledge acquisition to accelerate R&D is increasing. This chapter explains how artificial intelligence can contribute to knowledge acquisition. The process of transforming information into knowledge can be divided into two stages: "curation" and "knowledge acquisition." AI supports both stages by integrating information, assigning meaning, and facilitating researchers' access to this knowledge. The scientific search system includes components developed in collaboration with our clients and is currently being used in real-world applications.

OUTLINE

Enthought powers digital transformation for science.
We partner with companies worldwide to solve complex data challenges unique to enterprise scientific R&D. By leveraging advanced technologies, we accelerate innovation and drive business transformation. We bring an unparalleled blend of expertise and experience in advanced AI/ML techniques, scientific research and data, and leveraging R&D to support the business. Enthought is headquartered in Austin, Texas, USA, with additional offices in Tokyo, Japan; Cambridge, United Kingdom; and Zürich, Switzerland.

Software Development & Implementation	Roadmap Strategy & Process Optimization	In-House Digital Capability Building

enthought.jp

Section 4: Artificial intelligence for curation of information and knowledge acquisition

Christopher L. Farrow, PhD and Alexandre Chabot-Leclerc, PhD
Enthought. G.K.

Introduction

For decades, digital technologies have played a growing role in materials research and product development innovation. The increase in demand for digital solutions for materials challenges is driven by market pressures to create specialized materials faster, such as for semiconductor manufacturing, or to create entirely new products for the betterment of society, such as batteries, materials and processes for utilization of greenhouse gasses, and sustainable plastics. Fundamental to research progress is the curation and acquisition of knowledge. Today, researchers employ a variety of technologies to help in this task, such as public and private document servers, databaases of domain-specific information sources, and online search services. These technologies have aided knowledge acquisition by expanding the amount of information available to researchers, improving access to information sources, and making it easier to find specific information. However, technology has yet to transform how researchers reason about and apply information. Technologies such as Generative AI can potentially create that transformation.

To understand how technology impacts knowledge acquisition, consider the process through which information becomes knowledge, as shown in Fig. 1. This process is broken up into two broad stages: Curation and Knowledge Acquisition . In the curation stage, information is collected and made ready for application to a particular task. This involves finding information sources, extracting and saving specific information that is deemed relevant, and then later (or immediately) retrieving it for use. Knowledge acquisition occurs when that information is re-interpreted within a specific context for a particular purpose. In other words, when a researcher has gathered information and draws new conclusions, the information becomes useful and is turned into knowledge.

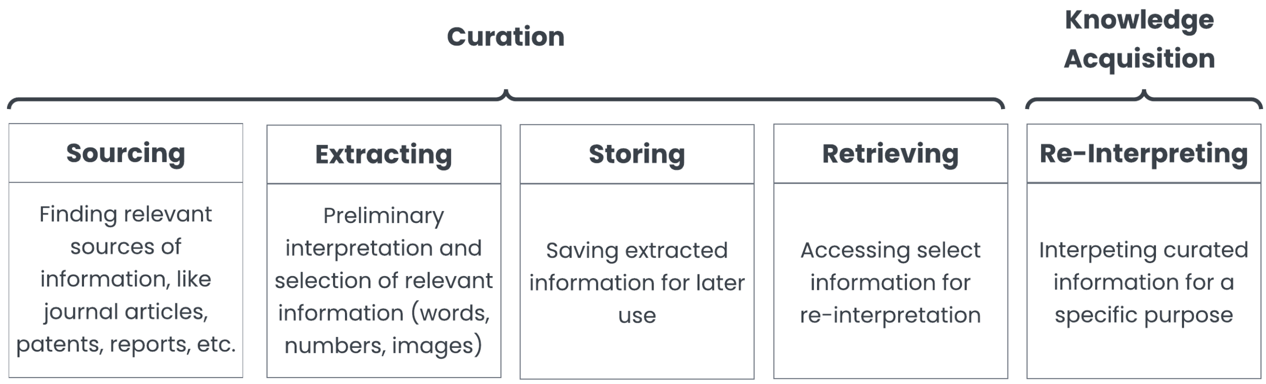


Figure 1. The process of curating information and acquiring knowledge

Many technologies and approaches are available today to simplify information curation, with search being the most notable. This will be discussed in the first part of this manuscript. Knowledge acquisition, however, is a technology frontier. Re-interpretation of information to create knowledge requires existing knowledge, as well as intuition and creativity—these are all human qualities. Artificial intelligence can mimic these qualities and support both curation and knowledge acquisition by fusing information, making semantic connections, and making those accessible to researchers. This will be discussed in the second and third parts of the manuscript.

1. Technology-Assisted Curation

Curation, when performed manually by people, is optimized to select relevant information as early as possible to make it manageable. Information selection is often done with a specific purpose, influencing the information that is extracted and saved. Indeed, researchers often interpret information and build knowledge as they are sourcing the information. They do not bother to store the extracted information for future use, as this is a cumbersome manual process.

It is important to note that information selection is information removal excluded from the curated set is not readily available for re-interpretation. This exclusion of information is inherent to many research processes. For example, researchers often discard images and spectra from laboratory analysis in favor of more manageable summary metrics that can be organized into tables. When those tables do not explain observed phenomena, the opportunity for knowledge acquisition is hindered.

To optimize knowledge acquisition, curation must be performed so that information is not permanently discarded. This way, it remains available to re-interpret in new contexts. Reviewing the curation process in Fig. 1, this means that the optimal technology-assisted curation system sources, extracts, and stores as much information as possible and eliminates irrelevant information only during retrieval. From this perspective, the optimal technology-assisted curation system is a search engine.

1.1 Curation as Search

Search engines can, in principle, serve up any digitized information. There are physical limits on storage for digital information and on the speed at which information can be retrieved, but available technology is sufficient for most practical purposes. Despite this, search engines typically deal with text—they respond to a text-based query by returning text and a link to the source of that text. Some search engines work with images and accept other images or text in a query. However, none provide comprehensive access to text, images, tables, graphs, formulas, molecular strings, and other information scientists must study to build knowledge. As a result, a researcher may perform many searches and scan many documents to extract information the search engine does not know how to handle, which amounts to manually curating non-text information.

The next step beyond text-based search is to incorporate scientific information into search queries and search results to ease or eliminate the need to study each table or graph to determine whether it is relevant. This requires a search engine that can handle this information and do so in a context-sensitive way. This context awareness is critical in research. To a search engine, nearly identical charge-discharge curves carry the same significance. To a researcher, the charge-discharge curve characterizing a material may be much more relevant to their research.

Developing a comprehensive search system that works in all contexts is infeasible, at least with traditional technologies. However, these technologies can be applied in new ways to greatly accelerate research and knowledge-driven workflows in contexts where the manual curation process is constrained and well understood. Enthought has built search systems like this for various knowledge-driven workflows. These are here referred to as scientific search systems because they enable researchers to search for or with scientific information. Fig. 2 shows a schematic of the generalized scientific search system. Enhanced search features of this system are described below.

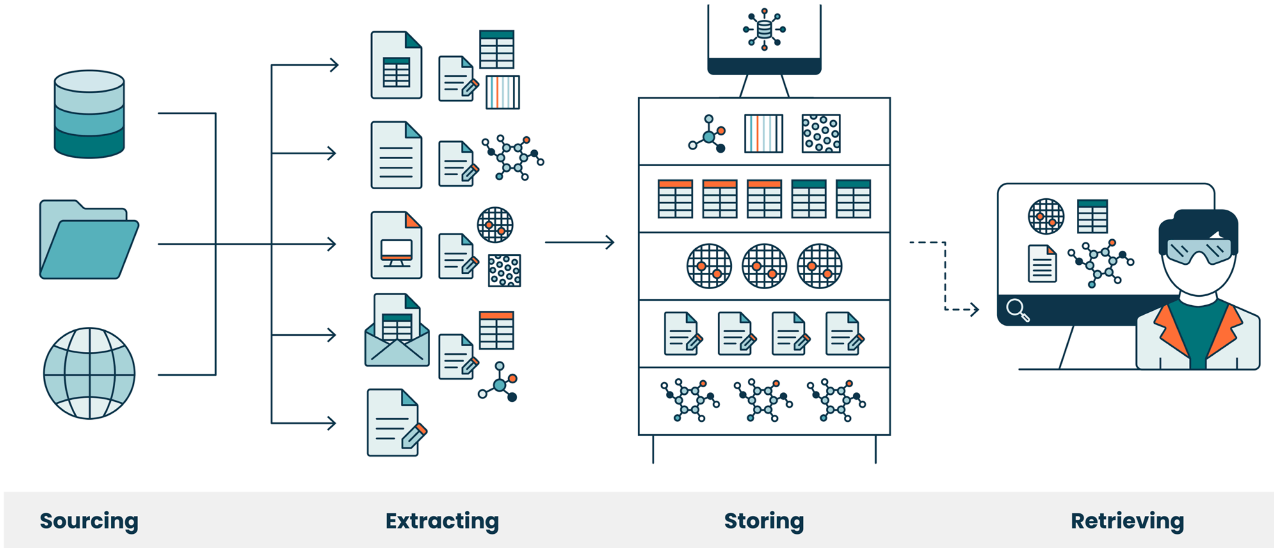


Figure 2: The generalized scientific search system. Information sources of various formats are deconstructed into text and other searchable content, like images, tables, and molecular formulas. All searchable information is categorized by type, with differentiation based on image or table content. The application interface allows researchers to input text and non-text queries to retrieve specific types of information.

1.1.1 NLP-enhanced search

Text-based search remains essential in the scientific search system, though it is enhanced by natural language processing (NLP). Most search engines flexibly interpret input text by removing tense, checking synonyms, and logically concatenating search terms by default. What few can do is interpret a query to return results that match intent rather than just words.

Leveraging neural networks and text embeddings gets one step closer to this. Modern research into natural language processing has produced various neural network architectures and pre-trained neural networks that represent words or phrases as a vector in a vector space, also known as text embedding. This vector space is learned from a corpus of text and has the characteristic that words and phrases that are similar have a short distance in this vector space.

Using text embeddings for text search amounts to finding words or phrases from the information sources that are within a specified cutoff distance from an input query. Searching by proximity within this vector space yields words and phrases that are not only similar in usage, but also similar in concept to an input query. For example, searching for the term 'redox' could return hits for 'electron transfer' (see Fig. 3). This use of word embeddings is closely related to Large Language Models, which are discussed more later in this manuscript.

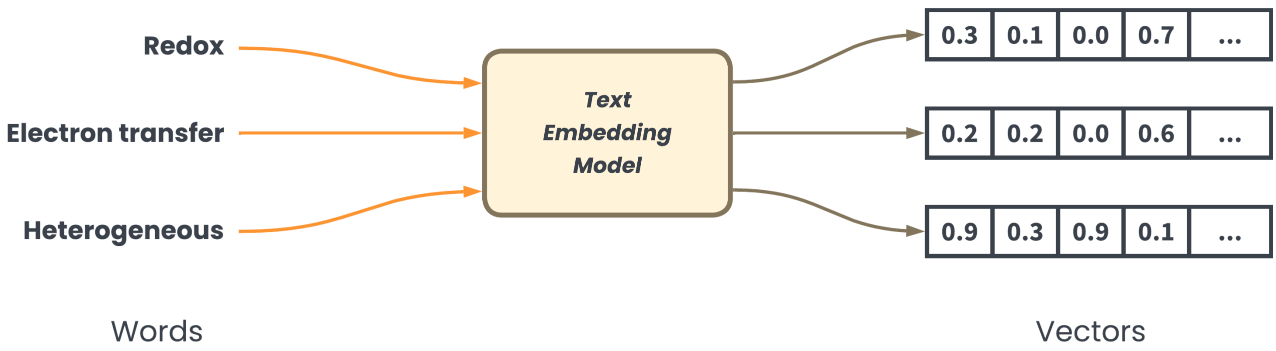


Figure 3: Embedding models map words and phrases to a vector in a high-dimensional space. Similar words and concepts, determined by how words and phrases are used in the training corpus, are nearby within the embedding space.

1.1.2 Image Search

Image-based search is widely available for finding images on the internet. One can go online now and type text or upload a picture to search for images of pink elephants or microchips. This uses neural network embeddings and similarity metrics, similar to the text embeddings described in the previous section. Generally, these online search engines have not been trained to identify or understand the differences a researcher would see in scientific images or graphs. For example, searching for 'charge-discharge curves for solid electrolyte batteries with graphene anodes' would return imprecise and in some cases irrelevant results.

The scientific search system in Fig. 2 improves upon generic image-based search by recognizing predefined classes of images and searching in-class whenever an image in that class is provided. For example, a researcher can upload an SEM image of nanoscale pillars etched onto silicon, and the system would only return similar images within its class. In the scientific search case, a limited and focused data set is an advantage because it represents fewer image classes than the variety of images available to a search engine operating at the scale of the world wide web. Searching for etched pillars and receiving pictures of cityscapes would be impossible.

The scientific search system requires each class of images to have its own 'detector,' a custom-trained neural network that can differentiate in-class images from out-of-class images. This architecture also allows a detector to extract class-specific information from retrieved images, which can reduce the effort required for preliminary interpretation of the image. The accuracy and sensitivity of these detectors can be enhanced by using transfer learning, which fine-tunes a general neural network to identify specific features within a smaller data set. This means the system can return reasonable results with less training data. To make the system even more robust, user-supplied ratings on search results are used to fine-tune the detectors when new information sources are added.

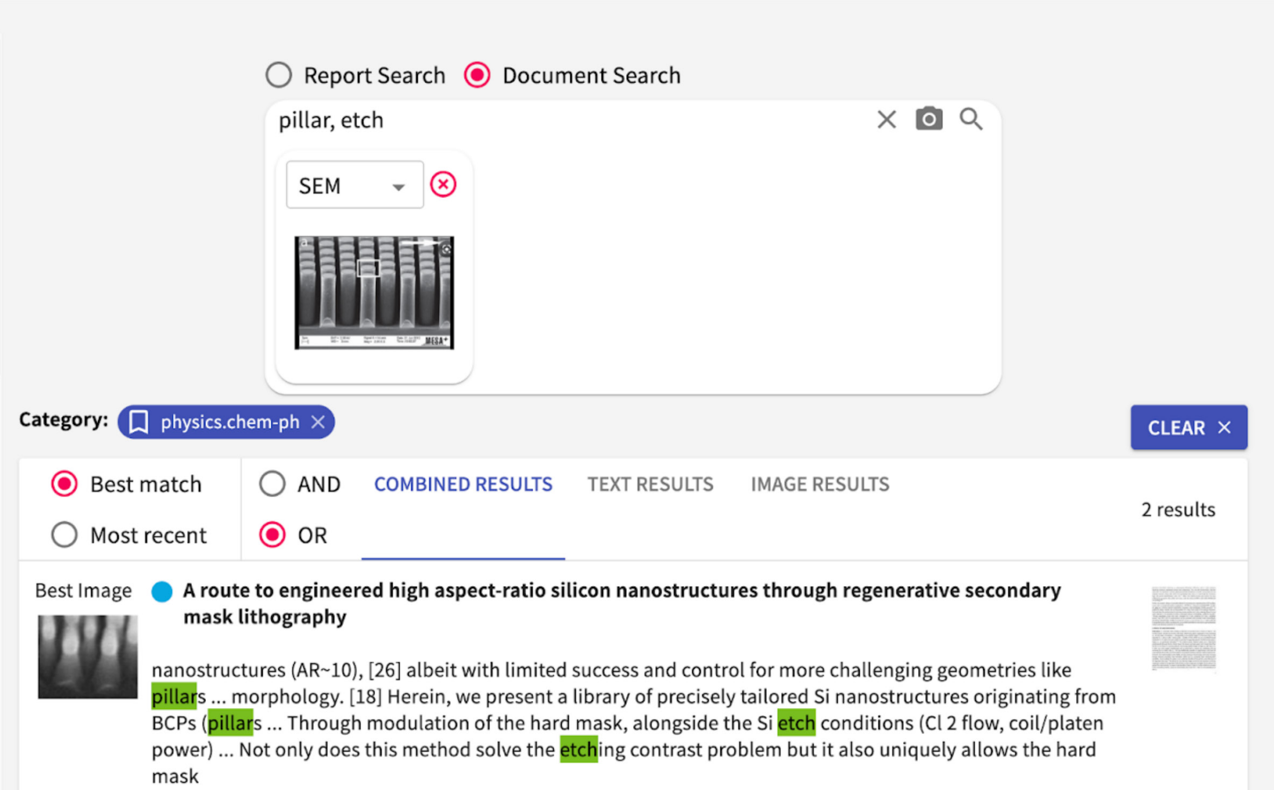


Figure 4: Scientific search application with image searching capability. The search terms 'pillar' and 'etch' are combined with an image to locate articles with similar images and terms. (Thank you to arXiv for the use of its open-access interoperability.)

1.1.3 Table Search and Domain-Specific Search

The table search functionality allows a researcher to find tables based on another table input as comma-separated values. This is not as sophisticated as image search, as it uses classical algorithmic methods to identify tables based on shape and headings rather than their visual appearance. Nonetheless, the functionality allows a researcher to locate relevant information sources faster than from text-based search alone.

The same scheme can be used to search for any specialized content that can be entered as text. This is implemented by providing special input boxes for these queries or by intercepting text input and applying rules to identify what the text represents. For example, a particular scientific search application enables molecular search by recognizing SMILES strings appropriately, and broadening the search to multiple molecular representations.

1.1.4 Extracting Data from Graphs

Finding and locating scientific information intuitively, combined with human-level image recognition, enhances the search experience. This still leaves to the researcher the work of interpreting and extracting data from graphs. The next-generation scientific search applications will be able to automate some of this, using a scheme similar to image search. Detectors are used to identify specific classes of graphs, each associated with algorithms for extracting key information. These extraction algorithms use image segmentation to identify specific graph features, such as axes titles, markers, labels, legends, and individual curves. From a segmented curve and this other information, each curve can be digitized with proper physical coordinates and units. The physical data extracted from the graph can be analyzed in arbitrary ways to aid research.

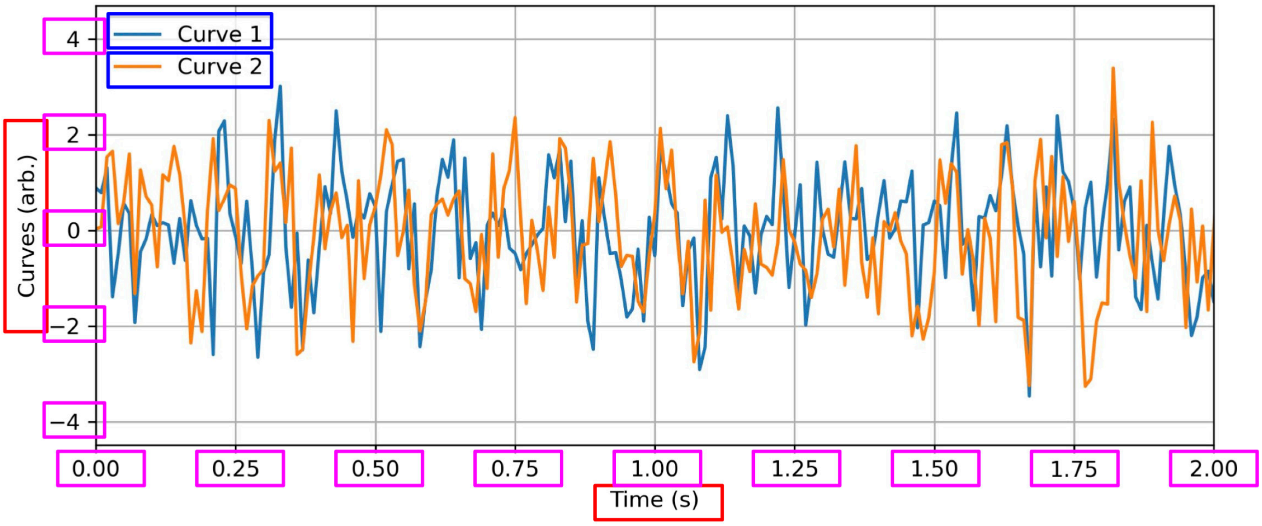


Figure 5: **An appropriately trained segmentation model can identify axes labels, tick marks, legends, and curves in graphs. The parsed segments provide enough information to extract data coordinates without manual work.**

1.2 Limitations of Search

Scientific search systems have successfully transformed how researchers curate and access information, and there is plenty more that can be done without modifying the scheme in Fig. 2. That said, search has limitations that cannot be addressed with this architecture. These limitations stem from inherent tradeoffs among how easy it is to use such a system, the quality of results it can provide, and the comprehensiveness of those results.

Google was the big winner in the early days of internet search due to its PageRank algorithm[1]. This algorithm worked by weighting the relevance of search results based on the number of other web pages that linked to each result. This was done without any understanding of the page contents, just connections among pages. The result was a higher quality of search results that was unmatched by other search providers. On top of this, Google was easier to use than other search engines because it interpreted user input more intuitively.

Similar connections exist in the form of citations across academic papers, a primary source for scientific research. However, industrial research typically requires proprietary information and other sources that may not provide explicit links to related information. This makes it difficult or impossible to automatically judge relevance in search and, therefore, to control the quality of information coming out of a search engine. Consequently, as more sources of information are made available to search, increasing the comprehensiveness of the system, the quality of search results will diminish.

One way to deal with a lack of connectivity is to allow more expressiveness in the search queries and more ways to search. Many search engines provide specialized search query languages and various search options for this purpose. While this may allow researchers to locate specific information with high precision, it also adds a user-experience barrier to improving the quality of search results.

Furthermore, even if such links existed across all information sources and could be easily leveraged for search, they may not be particularly useful. Scientific discovery is often hidden amongst the discoveries and concepts that have yet to be explicitly connected. References do not necessarily elicit those connections, and reference counts may not correlate with relevance for a particular purpose.

These search limitations stem largely from how information is shared and the search technologies used to retrieve it. Large Language Models and Generative AI provide new options for search and overcome many of these limitations. This has implications for both curation and re-interpretation of information, which are explored in the following sections.

2. Generative AI for Curation

Advances in deep learning technologies brought forth by the Transformer network architecture[2] affect every stage of the curation process. So-called Large Language Models (LLM) are the most well-known use of transformers, with models such as OpenAI's GPT-3[3] and GPT-4[4], Anthropic's Claude[5], and Meta's Llama 2[6]. The other prominent use of transformers is for text-to-image creation, such as OpenAI's DALL-E[7], Stability AI's Stable Diffusion[8], and Midjourney[9]. In addition to these two use cases, many models have been applied to other use cases, such as image-to-text, image segmentation, image labeling, document understanding, and more. Collectively, these models have been termed Generative AI .

The power of Generative AI comes from how it learns and generates information. In the case of LLMs, the learning amounts to fitting the parameters of a statistical model to training data, which is ingested as tokens . When the model is used to generate new tokens, patterns that are prevalent in the training set tend to re-emerge in ways that resemble the training data. This can be used to summarize information that is spread across multiple sources. When a model incorporates multiple types of information, say text and images, this generative can be used to transform one type to another. This greatly expands how researchers can find and access multiple types of information.

2.1 Everything Can Become Text

In general, search engines are text-centric. Though it is possible to search with and retrieve non-text information, text-based context is often needed to interpret the information. Primary and contextual text comes from a variety of sources—text documents, text extracted from images with optical character recognition (OCR), audio converted to text, tables, captions, or manually-supplied labels—but it is not always easy to obtain text that is useful.

Linking captions to images is relatively simple in structured text, such as web documents. Unstructured or semi-structured text documents provide unique challenges. PDF documents generated using particular tools can provide structure for associating captions and images, whereas a PowerPoint presentation may have no reliable structure that links images and captions. Indeed, not every image will have a caption or associated text. Manually labeling images is one way to enable text-based search of images, but this is a tedious and expensive task that requires expertise and knowledge of how downstream users may query for this image. Labeling videos is even more expensive, and is rarely done. Challenges like these have limited the richness of information that can be retrieved through text-based search.

Generative AI can help address this. A family of Transformer models enables the conversion of various content to text, alleviating this labeling challenge and unlocking new querying possibilities with text- based search tools. For example, BLIP[10] and GIT[11] are image auto-captioning models that can describe the content of images. The descriptions go beyond simple labeling of objects in an image, and include relationships and actions among the objects. Similarly, models such as SwinBERT[12] and VALOR[13] go beyond converting speech to text and describe what is happening in videos.

For scanned documents, where text can only be interpreted visually, another family of models offers OCR-free document understanding. The approach sidesteps a major challenge of OCR, which is that character recognition errors degrade subsequent interpretation of document content. By looking at an entire document, or recognizable units, models such as Donut[14] and Pix2Struct[15] understand and extract text and data without interpreting characters individually. This provides a promising avenue for deeper interaction with scanned documents or document images, such as data extraction.

2.2 Text Can Become Data

Text documents often contain critical numerical or relational data that are essential to answering a research question. Once again, extracting that data from text is manual, time-consuming, and requires expertise. Even when the data is already in tabular format, queries and extraction can be challenging, as discussed in 1.1.4. For example, answering the question 'What is the record for highest temperature at which superconductivity has been observed and verified?' is easy if the data is available in a database—a simple SQL query—but may require reading multiple documents before finding the one with the correct and latest information.

LLMs can bridge the gap between unstructured and structured data and can answer such a query on demand. In more complex cases, data can be extracted from prose using a fine-tuned LLM[16] or a chat interface using appropriate prompts[17]. This approach can be applied to a whole list of documents retrieved from a keyword or semantic prompt, yielding an array of structured data. If the goal is to store information for future use, these tools can be used to more easily build databases of numerical data with links to sources. Ideally, within an organization, raw and processed data is accessible according to some rational and structured scheme. But, in the (frequent) cases where researchers exchange information using reports and presentations, LLMs may help restore historical data to answer new questions.

2.3 Beyond Text Search: Multi-Modal Embeddings

Section 1.1.1 mentions text embeddings as the critical ingredient for semantic search : finding chunks of text that share the same meaning as the query, even if they do not share keywords. This concept has recently been extended to support multimodal embeddings, i.e., embeddings of multiple data types within the same high-dimensional space. These embeddings enable cross-modality querying without going through a shared medium like text.

As an example, ImageBind18 embeds six media types into the same latent space: images, text, audio, depth, thermal, and inertial measurement unit (IMU) data. This model makes it possible to surface relevant images given an audio query or relevant depth maps given a text query. One can imagine extending this capability to the laboratory to predict properties from indirectly related measurements. For example, one could input an SEM image of a solid state electrolyte to predict the ionic conductivity of that sample. The challenge of building a model like this would be acquiring enough training data to produce useful results.

A powerful aspect of these image embeddings is that they encode not only the category of images but also the identity and properties of their components. This means search engines can retrieve images based on their content, not just their class. Both ImageBind and another image retrieval model, TASK- former[19], enable the arithmetic combination of modalities to create a query. This enables searching based on coincidence of modalities, such as a specific combination of text and images. This is akin to providing additional keywords to narrow down a search, extended to non-text information.

3. Generative AI for Knowledge Acquisition

So far, computing has contributed to every stage of the information curation process, but has been mostly absent from knowledge creation. One reason is that classical search engines do not answer queries, even simple ones like 'Is titanium heavier than aluminum?' A search engine may return a series of links to documents, and sometimes snippets, but not an answer. It is up to a researcher to read the documents, to check whether they are relevant, and reason about the information to answer the research question. More complex queries, such as 'What are the top 5 advances in battery technology in the last five years?' are completely out of reach for a classical search engine, unless a document already exists on this topic.

Because Large Language Models synthesize information from multiple sources, as described in the previous section, they can generate answers to questions like these. In doing so, the model effectively decides what is relevant and what is not. This is done through token-level learning of the information, not true intelligence, so it does not completely free researchers from the work of building knowledge. Nonetheless, LLMs and other Generative AI can serve as competent research assistants (that sometimes make mistakes). This section describes how those assistants can be leveraged to answer questions, perform work, and make connections a human researcher may miss.

3.1 Retrieval-Augmented Generation for Answering Questions

A core pattern required for LLMs to answer research questions, rather than just returning links to documents, is called retrieval augmented generation (RAG)[20]. RAG aims to enable LLMs to act and rely on up-to-date and correct information without retraining the model as new information comes in. It pairs a search engine with an LLM, which then can serve as a 'thinking' and summarization tool, rather than just a knowledge base. First, the search engine retrieves documents based on a researcher's query. Second, the records retrieved are concatenated once together, and again with the researcher's query. Third, this whole text is passed to an LLM to answer the query given the context fetched from the documents. This is shown in Fig. 6. The user's query can be augmented with additional prompts or tooling to shape the response. For example, a common addition is to require the model to 'cite its sources' (from the context document) along with the answer.

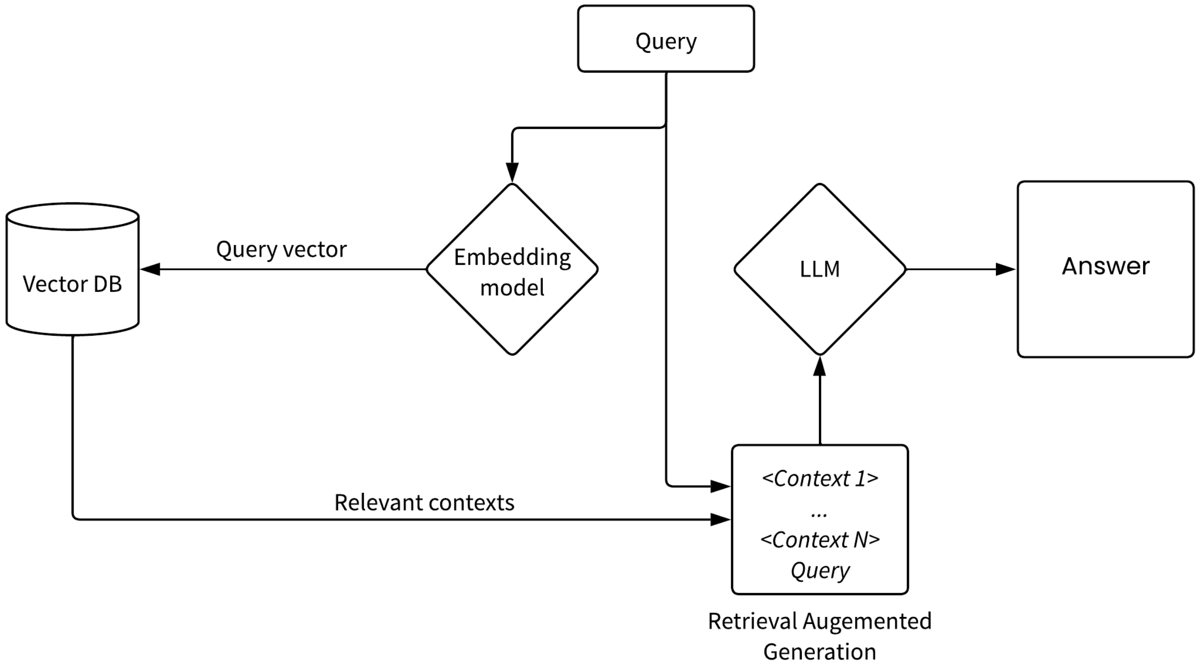


Figure 6: Architecture of the Retrieval Augmented Generation pattern[21]. Relevant documents are pulled from the vector database and concatenated together with the prompt. The result is sent to the LLM to answer the query.

The RAG pattern is technically simple but presents at least two challenges. The first is selecting the relevant documents to ensure the LLM can answer the question. This responsibility falls on the upstream search engine, whether it uses keyword search, embedding search, or other techniques. The second challenge is the length of the documents retrieved. LLMs have a limited input size, or context length , expressed in tokens. A token represents a commonly occurring sequence of characters. One token typically represents four characters in English, yielding about 0.75 words per token. Most models have a context length of 4k to 8k tokens, with some models going up to 16k, 32k, and 100k. For reference, the graphene patent has about 6k tokens, whereas the insulin pump patent has 48k. Until larger context sizes are widely available, documents must often be summarized to fit within the context length.

Many transformer models can perform text summarization. Some of these are specialized models that are trained to only perform text summarization[22]. General-purpose models can also be tuned to follow instructions[23-24] to summarize text when prompted to do so: 'Provide a clear and concise summary of this text: <text>.' Most models perform abstractive summarization , where the model generates an original document summary based on its semantic representation. This type of summarization is akin to how a human summarizes a document. This can be applied to summarize individual documents, helping a researcher screen whether or not to read them. Alternatively, a researcher can get a broad picture of a whole section of literature by summarizing summaries from multiple documents, or by using those summaries in RAG to make specific queries.

Combining summarization with data extraction can provide an even richer picture from a collection of documents. For example, the Elicit search engine[25] surfaces research papers and summarized abstracts relevant to a query. It allows users to ask the same ‘question’ from each paper and displays the answer as an additional column. This can be used to surface something like sample size or a measured outcome from multiple articles, freeing the researcher from skimming or reading each paper to help select the most relevant one.

Once a researcher has identified a relevant paper, LLMs can provide additional support while investigating the document. One pattern is to give a now-familiar chat interface to the user so they can ask questions about the document. This question-and-answer pattern is also implemented as RAG. Various services and products are being developed to provide interfaces beyond chat to enrich the reading experience. For example, ExplainPaper[26] lets the user highlight part of the document and ask for an explanation of the highlighted concept. The user can even tune the level of complexity of the explanation.

3.2 Agents for Doing Work

At the frontier of search and document retrieval with LLMs and embeddings are agents : LLMs that make plans[27] and use tools that allow them to do things beyond generating text. OpenAI’s ChatGPT with GPT-4 provides an ‘Advanced Data Analysis’ mode that can write code, analyze data, and create figures without access to the internet. The ChemCrow[28] project is a domain-specific agent that can answer questions beyond requests for facts. In ChemCrow, an LLM is embedded in an iterative chain-of-thought process[29] that lets the model plan its actions, select a tool, use it, observe the result, and repeat the planning phase until the task or query is complete. Given access to 13 tools, such as fetching SMILES strings for a molecule, accessing a molecular synthesis planner, and accessing safety information databases, ChemCrow could complete the following tasks: ‘I need to synthesize a sample of atorvastatin. Please tell me how to synthesize it. Then tell me how much it will cost to buy all the reactants I need. Provide specific substance names (iupac) or smiles. If you can’t determine the price of the reactants, try another synthetic route.’[28]

Pairing an LLM’s ability to write code and to execute it with tools, the LLM can act as a translator between ‘plain language’ queries and a structured query to fetch data from structured databases. For example, it can generate SQL and execute it against a relational database. Alternatively, an LLM could leverage the growing list of knowledge graphs, such as System[30] and DiffBot[31], to help itself answer a question. One way to think about it is that, given access to tools, LLMs can retrieve the information they need to add to their context to answer the user's query.

The advancements at the querying and knowledge acquisition stage will require significant design, user interface, and usability improvements to reach their potential. A free-form text box where users type their query provides the ultimate flexibility, but little discoverability about what is possible. Some tasks can be expressed succinctly via text, while others benefit from direct user interactions. For example, it is easier and a richer experience to extract information from a video using fine-control sliders with immediate feedback rather than asking the model to 'transcribe reaction equations shown on screen between 2:25 to 2:46' and following up with 'actually check until 2:55.' It's a salient example of the limitations of text input when a more direct interaction model would be better. Given a task, an exciting potential is for the LLM to generate the appropriate user interface on the fly. Though the authors could not find examples of this idea at the time of this writing, there has been an investigation into using one LLM to generate tools for another to solve a task[32]. The result suggests it is feasible for LLMs to generate user interfaces for human researchers.

3.3 Embeddings for Making Connections

By encoding training data in a dense embedding space, LLMs connect otherwise disparate information at a semantic level. This can be thought of as packing information with similar meaning into nearby or overlapping areas of a high-dimensional space. By navigating that space, the latent connections among information can be used for idea generation and learning.

One way to approach this is to use the t-distributed stochastic neighbor embedding technique (t-SNE[33]), which generates a 2D image that faithfully represents clusters in the original high- dimensional embedding space. Going from one point in this space to another nearby can take you to a related concept or representation. Equipped with an interactive image, and links from points to information sources, a researcher could easily explore related ideas and concepts within a training set, though in a somewhat undirected way.

Another approach, using an LLM, is much simpler. A researcher can simply ask an LLM to make connections to an idea at hand, and the latent connections will manifest in the response it generates. An example is shown in Fig. 7.

LLMs have proven to be good at this sort of creative generation. Yet, the user has to be careful about the model's tendency to hallucinate, i.e., to provide a confident response despite it being wrong and unjustified based on its training data[34]. The quality of the response will also be dependent on the data the model has been trained on, with the same limitations leading to the creation of RAGs: the model's 'knowledge' is stuck at a particular point in time and doesn't have access to up-to-date information unless it's provided to it as part of the prompt's context.

User: Chemistry has this concept of 'activation energy.' What ideas from other fields are either similar or related to this concept?

Answer: There are several concepts from other fields that are similar or related to the concept of activation energy in chemistry. Here are a few examples:

1. Physics - Potential Energy Barrier: The concept of potential energy barrier in physics is similar to activation energy in chemistry. It refers to the minimum energy required to overcome a barrier or reach a new state in a physical system. This concept is often used to explain phenomena like the activation of particles in nuclear reactions or the transition of electrons between energy levels.

2. Psychology - Threshold of Activation: In psychology, the concept of threshold of activation is closely related to activation energy. It refers to the minimum level of stimulation or intensity required to trigger a response or initiate a particular behavior. Just like in chemistry, where a certain energy is needed to start a reaction, psychology recognizes that a certain level of stimulus is necessary to activate a response or behavior.

Answer, continued:

3. Biology - Threshold of Excitation: In the context of neurons and nerve impulses, the concept of threshold of excitation is related to activation energy. It represents the minimum level of depolarization required to initiate an action potential or nerve impulse. Similar to activation energy in chemistry, exceeding this threshold is necessary for the nerve impulse to propagate.

4. Economics - Barrier to Entry: In economics, the concept of barrier to entry is comparable to activation energy. It refers to the obstacles or costs that new entrants face when trying to enter a particular industry or market. Just as activation energy determines the feasibility of a chemical reaction, barriers to entry can limit the entry of new competitors and affect the dynamics of an industry.

5. Sociology - Tipping Point: In sociology, the concept of a tipping point is similar to activation energy. It represents the critical threshold or level of adoption required for an idea, behavior, or trend to spread rapidly within a social system. Like activation energy, reaching the tipping point is essential for initiating a large-scale change or transformation in society.

These concepts from various fields share similarities with activation energy as they all involve a minimum threshold or barrier that needs to be surpassed to trigger a change, response, or transition in a system.

Figure 7: Example of a conversation between a researcher and an LLM (ChatGPT, asking the LLM for concepts relating to 'activation energy.'[35])

Conclusion

Large Language Models and Generative AI are already changing how researchers acquire knowledge. Previous generations of technologies can be leveraged in many ways to assist in the curation of information. Search is the optimal scheme for curation, as it allows for selection of relevant information once an intended use of that information is known. The research utility of search can be further enhanced by incorporating non-text information, such as images, tables, formulas, graphs, and so on. This is not without limitations, however, as the lack of semantic connections among information sources leads to degraded search quality as more information sources are added. This limitation is very difficult to overcome, as the semantic connection that maximizes the relevance of search results is dependent on the intent of the search, which in general cannot be identified or easily made explicit when building and indexing a corpus to search.

Generative AI and LLMs have taken a large step in overcoming this limitation. These technologies make semantic connections based on how information is presented in text and other forms. The information that feeds these models is largely created for human consumption, and as such the intent and semantics of the source material can be more faithfully captured and utilized for search. Consequently, LLMs can provide an intuitive, natural-language interface to information.

The now-familiar chat interfaces to LLMs are just a starting point for how researchers will eventually interact with Generative AI. Already, Generative AI can answer questions, do work, and act as a research assistant that contributes to creative work. There is much room to expand the functionality of these new research tools, and improve how we interact with them, but in this first year of LLMs and ChatGPT entering public awareness, the innovations are already astounding. While public expectations of what is possible with Generative AI will be tempered in the coming years, the transformation of research and knowledge acquisition brought about by these technologies, some which are predicted here, will certainly continue.

Acknowledgment

Special thanks to the Enthought team, whose work to empower industry researchers through digital innovation produced the scientific search system described here.

Portions of the scientific search system were developed in collaboration with Tokyo Electron Ltd. and are used in Tokyo Electron Ltd.

References

1. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The pagerank citation ranking: Bring order to the web." *Technical report, Stanford University* (1998).
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances In Neural Information Processing Systems* 30 (2017). https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Ab- stract.html
3. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *Advances In Neural Information Processing Systems* 33 (2020): 1877-1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Ab- stract.html
4. OpenAI. "GPT-4 Technical Report." ArXiv abs/2303.08774 (2023) https://arxiv.org/abs/2303.08774
5. Anthropic, "Claude 2," July 11, 2023, Accessed on September 21, 2023, https://www.anthropic.com/index/claude-2
6. Meta, "Llama 2: Open Foundation and Fine-Tuned Chat Models," July 18, 2023, Accessed on September 21, 2023, https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/
7. OpenAI, "DALL·E: Creating images from text," January 5, 2021, Accessed on September 21, 2023, https://openai.com/research/dall-e
8. Stability AI, "Stable Diffusion Launch Announcement," August 10, 2022, Accessed on September 21, 2023, https://stability.ai/blog/stable-diffusion-announcement
9. Midjourney, "Midjourney," Accessed on September 21, 2023, https://www.midjourney.com/home/
10. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *International Conference on Machine Learning*, pp. 12888-1290,. PMLR, (2022) https://proceedings.mlr.press/v162/li22n.html

11 ianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang, "GIT: A generative image-to-text transformer for vision and language." *ArXiv preprint arXiv:2205.14100* (2022) https://arxiv.org/abs/2205.14100

12. Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang, "SwinBERT: End-to-end transformers with sparse attention for video captioning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17949-17958, (2022) https://openaccess.thecvf.com/content/CVPR2022/html/Lin_SwinBERT_End-to-End_Transformers_ With_Sparse_Attention_for_Video_Captioning_CVPR_2022_paper.html
13. Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu, "VAL- OR: Vision-audio-language omni-perception pretraining model and dataset." *ArXiv preprint arXiv:2304.08345* (2023) https://arxiv.org/abs/2304.08345
14. Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Won- seok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park, "OCR-free document understanding transformer," *European Conference on Computer Vision*, pp. 498-517, Cham: Springer Nature Switzer- land (2022) https://link.springer.com/chapter/10.1007/978-3-031-19815-1_29
15. Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova, "Pix2struct: Screenshot parsing as pretraining for visual language understanding," *International Conference on Machine Learning*, pp. 18893-18912, PMLR (2023) https://arxiv.org/abs/2210.03347
16. Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain, "Structured information extraction from complex scientific text with fine-tuned large language models," *ArXiv preprint arXiv:2212.05238* (2022) https://arxiv.org/abs/2212.05238
17. Maciej P. Polak and Dane Morgan, "Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering," arXiv preprint arXiv:2303.05352 (2023) https://arxiv.org/abs/2303.05352
18. Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra, "Imagebind: One embedding space to bind them all," *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180-15190 (2023) https://arxiv.org/abs/2305.05665
19. Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays, "A sketch is worth a thousand words: Image retrieval with text and sketch," *European Conference on Computer Vision*, pp. 251-267. Cham: Springer Nature Switzerland (2022) https://arxiv.org/abs/2208.03354
20. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel,

and Douwe Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems* 33 (2020). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
21. Pinecone, “LLMs Are Not All You Need,” September 11, 2023, Accessed on September 21, 2023,
https://www.pinecone.io/learn/llm-ecosystem/
22. Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” *International Conference on Machine Learning*, pp. 11328-11339, PMLR (2020)
https://proceedings.mlr.press/v119/zhang20ae.html
23. Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems* 35 (2022)
https://arxiv.org/abs/2203.02155
24. OpenAI, “Aligning language models to follow instructions,” January 27, 2022, Accessed on September 21, 2023, https://openai.com/research/instruction-following
25. Elicit. “Elicit: The AI Research Assistant,” Accessed on September 21, 2023, https://elicit.com/
26. Explainpaper. “Explainpaper,” Accessed on September 21, 2023,
https://www.explainpaper.com
27. Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch, “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents,” *International Conference on Machine Learning*, pp. 9118-9147, PMLR (2022)
https://proceedings.mlr.press/v162/huang22a.html
28. Andres M. Bran, Sam Cox, Andrew D. White, and Philippe Schwaller, “ChemCrow: Augmenting large-language models with chemistry tools,” *arXiv preprint arXiv:2304.05376* (2023) https://arxiv.org/abs/2304.05376
29. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems* 35, 24824-24837, (2022)
https://arxiv.org/abs/2201.11903
30. System Inc. “System: Overview,” Accessed on September 21, 2023, https://about.system.com/product/system
31. DiffBot. “Diffbot | Knowledge Graph, AI Web Data Extraction and Crawling,” Accessed on September 22, 2023,
https://www.diffbot.com/

32. Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou, “Large language models as tool makers,” *arXiv preprint arXiv:2305.17126* (2023), https://arxiv.org/abs/2305.17126
33. Laurens Van der Maaten and Geoffrey Hinton, “Visualizing data using t-SNE,” Journal of Machine Learning Research 9, no. 11 (2008), https://www.jmlr.org/papers/v9/vandermaaten08a.html
34. Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys* 55, no. 12, 1-38, (2023), https://arxiv.org/abs/2202.03629
35. Open AI, “GPT-3.5 Turbo”, August 3, 2023, Accessed on September 21, 2023, https://chat.openai.com/

For more helpful resources,
visit enthought.jp

Enthought

Powering Digital Transformation for Science

Asia

Tokyo, Japan

Ark Hills Front Tower 16F, WAW Akasaka
2-23-1 Akasaka
Minato-ku, Tokyo
107-0052 Japan
+81 (0)3 6824 4578

North America

Austin, Texas (Headquarters)

Enthought
200 W Cesar Chavez, Suite 202
Austin, Texas 78701
United States
+1 512 536 1057

Europe

Cambridge, England

Enthought
Broers Building, 3rd Floor
21 JJ Thomson Avenue
Cambridge CB3 0FA
United Kingdom
+44 1223 321679

Zurich, Switzerland

Enthought
c/o Treureva AG
Othmarstrasse 8
8008 Zürich
Switzerland
+1 512 536 1057

enthought.jp
info@enthought.com

@enthoughtJapan

@enthought